**Data Mining Assignment 1: Statistical Analysis of Data**

**Due Date: 3rd October 2024 11:55 pm**

**Objective:**

The objective of this assignment is to analyze a synthetic dataset using Excel, applying various data mining techniques such as descriptive statistics, data visualization (including box plots and scatter plot matrices), and exploratory data analysis. You will develop insights based on your analysis and answer specific questions regarding the dataset.

**Dataset:**

You will work with the synthetic dataset generated in the previous task, which contains 50 samples with the following attributes:

* **3 Numerical Attributes**: Numerical1, Numerical2, Numerical3
* **3 Ordinal Attributes**: Ordinal1, Ordinal2, Ordinal3
* **3 Nominal Attributes**: Nominal1, Nominal2, Nominal3

**Assignment Tasks:**

1. **Data Preparation**:
   * Open the synthetic\_data.xlsx file.
   * Review the dataset to familiarize yourself with the structure and the types of attributes.
2. **Descriptive Statistics**:
   * Calculate the following summary statistics for the numerical attributes:
     + Mean
     + Median
     + Standard Deviation
     + Minimum
     + Maximum
   * Present your findings in a new worksheet called "Descriptive Statistics."
3. **Frequency Distribution**:
   * For each ordinal attribute, create a frequency distribution table that shows the count of each category.
   * For nominal attributes, create similar tables to count occurrences of each unique value.
   * Present these tables in a new worksheet called "Frequency Distribution."
4. **Data Visualization**:
   * Create the following visualizations:
     + **Box Plots**: Create box plots for each numerical attribute to show the distribution and identify potential outliers. Place these in a worksheet called "Box Plots."
     + **Scatter Plot Matrix**: Create a scatter plot matrix (pairwise scatter plots) for the three numerical attributes to visualize relationships between them. You can use a separate worksheet for this called "Scatter Plot Matrix."
     + **Pixel-Oriented Visualization**: Use conditional formatting to create a pixel-oriented heat map for one of the numerical attributes. For example, format Numerical1 in a heat map style to visualize its values. Place this in a worksheet called "Pixel Visualization."
     + **Additional Charts**: Include at least one pie chart for nominal data and a bar chart for ordinal data as previously described.
5. **Correlation Analysis**:
   * Calculate the correlation coefficients between all pairs of numerical attributes.
   * Present your findings in a new worksheet called "Correlation Analysis."
   * Highlight any strong correlations (close to 1 or -1).
6. **Pivot Table Analysis**:
   * Create a pivot table to analyze the relationship between one ordinal attribute and the average of one numerical attribute. For example, analyze the average Numerical1 by Ordinal1.
   * Present this pivot table in a new worksheet called "Pivot Table Analysis."
7. **Distance Matrix Calculation**:
   * Calculate the distance matrix for all attributes (numerical, ordinal, and nominal).
   * Use an appropriate distance metric for each attribute type discussed in class in Lecture 2
   * Present this combined distance matrix in a new worksheet called "Distance Matrix."
8. **Insights and Conclusions**:
   * Write a summary (1-2 paragraphs) in a new worksheet called "Insights" that discusses:
     + Key findings from your analysis.
     + Any surprising trends or relationships discovered in the data.
     + Recommendations for further analysis or actions based on your findings.

**Questions for Analysis:**

After completing the tasks, answer the following questions in the "Insights" worksheet:

1. **Descriptive Statistics**:
   * What is the average value of Numerical1? How does it compare to Numerical2 and Numerical3?
2. **Frequency Distribution**:
   * Which category in Ordinal1 has the highest frequency? What percentage of the total does this represent?
   * In the nominal attributes, which value in Nominal2 is the most common?
3. **Data Visualization**:
   * Based on the box plots, do you observe any outliers in any of the numerical attributes? What might this imply?
   * In the scatter plot matrix, describe any notable relationships between the numerical attributes.
   * What insights do you gain from the pixel-oriented heat map visualization?
4. **Correlation Analysis**:
   * Which pairs of numerical attributes have the strongest correlation? What does this imply about their relationship?
5. **Pivot Table Analysis**:
   * What is the average Numerical1 value for each category in Ordinal1? Are there significant differences among the categories?
6. **Overall Insights**:
   * Based on your analysis, what recommendations would you provide for further data collection or analysis? What limitations did you encounter during your analysis?

**Submission Requirements:**

* Submit your completed Excel file with all worksheets and analyses.
* Ensure that your insights are clear and well-supported by the data.
* Save your Excel file with a meaningful name, such as Data\_Mining\_Assignment\_YourName.xlsx.

**Evaluation Criteria:**

* Completeness of analysis and accuracy of calculations.
* Clarity and organization of the Excel file.
* Quality and depth of insights and conclusions drawn from the data.
* Presentation of visualizations and tables.

This assignment aims to provide hands-on experience with data mining techniques using Excel and to develop critical thinking skills when analyzing and interpreting data. If you have any questions or need further clarification, feel free to ask!